学校编码：10384　　密级_____

学号：X2006221020

厦门大学

工　程　硕　士　学　位　论　文

基于 Web 数据挖掘的研究及应用

The Research and Application Based on Web Data Mining

朱国文

指导教师姓名：　冯少荣　副教授
专　业　名　称：　计算机技术
论文提交日期：　2010 年 4 月
论文答辩日期：　2010 年 6 月

2010 年 6 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外,该学位论文为(　　)课题(组)的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

网站已成为 Internet 上主要的信息发布方式。如何利用 Web 站点现有的信息提高站点的性能和更好地为用户服务，已成为计算机应用领域的热门话题。Web 日志挖掘是 Web 数据挖掘的一个重要分支，它的目的在于从 Web 日志中挖掘网站的频繁使用模式、用户访问行为模式、具有相似行为的用户群等信息。通过 Web 日志挖掘能够充分了解 Web 站点的使用情况和使用 Web 站点的用户行为模式，从而改进 Web 站点性能，完善站点结构，更好地提供个性化服务，提高 Web 站点的访问量。

大多数 Web 数据挖掘系统无法与数据库无缝集成，也不支持算法的扩展。由于数据挖掘标准语言的欠缺，使 Web 数据挖掘技术的应用范围仅限于领域专家。而 SQL Server 2008 为数据挖掘解决方案提供了强大的设计和开发平台，方便了企业级的数据挖掘系统的设计和实现。

本文介绍了 Web 数据挖掘的基本概念及国内外研究现状，对 Web 日志进行了研究，提出了 Web 数据挖掘方法。针对《平和网》的 Web 日志，给出了一个基于 SQL Server 分析服务构建数据挖掘解决方案的方法，构造了相应的系统结构，设计并实现了一个 Web 数据挖掘系统。通过本系统，不仅可以获得《平和网》的基本统计信息，如站点的使用情况和服务器的响应情况，而且可以获得网站用户的访问模式和用户的聚类群信息。

关键词：Web 数据挖掘；SQL Server 2008；多维分析；聚类分析

Abstract

The Internet site has become the main ways to send information. How to use the Web site, to improve its performance and better serve its customers has become a hot topic of computer applications. Web log mining is an important branch of Web data mining. Its purpose is to find the modes of the frequent user, the patterns of users' behavior, and the similar actions of user groups in the Web log. Through a Web log mining we can fully understand the development and the model of users' behavior of Web site, the data of which will improve the function, the structure and personalized service of web sites. Therefore, the number of visitors to the web site can be increased.

Because most Web data mining systems and database do not support the seamless integration and algorithmic expansion. And the lack of standard language in data mining has made the application range of Web data mining technology limited to sphere experts, but SQL Server 2008 can provide a powerful design and development platform for the data mining solution and bring great convenience to the data mining system design and realization for the enterprise.

Through the study of the Web log, this paper introduces the basic concept of Web data mining and its research background both at home and abroad and put forward the methods of Web data mining. The Web log of *Ping-he Web* has played an important role in providing a method for solving the construction of data mining based on SQL Server, making up a corresponding system structure, designing and implementing a Web data mining system. Through this system, not only the basic statistical information of *Ping-he Web*, such as the developments of sites and responses of its server, but also the visiting mode of users and clustering information of users can be obtained.

Key Words: Web data mining; SQL Server 2008; Multidimensional analysis; Clustering analysis

目 录

Table of Contents

第一章　绪论

1.1　研究背景和意义

网站已成为 Internet 上主要的信息发布方式。不管是政府、企业还是个人，都开始采用网站的形式进行宣传和交流。作为网站运营者，不论是从构建网站布局也好，还是网页的内容也好，都希望给客户提供更多他们感兴趣的信息。从而获得访客更大的关注，那么每个人会有每个人查找信息的习惯，这就导致了他们浏览网页的顺序和方式会有所不同。对于如何提高网站的访问率，增强网站内容的时效性和针对性，以及如何更好地为用户提供个性化信息推荐服务，成为网站管理员关心的热点问题。

Web 日志挖掘就是解决这类问题的一种有效方法。Web 日志挖掘是 Web 挖掘的一种，指对用户访问 Web 时留下的访问记录进行挖掘。Web 日志体现了用户使用 Web 资源的行为特点以及隐藏在行为背后的更深层次的动因和规律。通过挖掘用户访问记录可以获得网站中的热点内容、相似用户群体和用户访问模式等信息；通过收集用户顺序请求的日期和时间，可以分析出用户在每个资源上所花费的时间，从而可以推断用户对该资源感兴趣的程度；通过收集用户感兴趣的领域，有利于对用户感兴趣的内容进行分类；通过分析用户请求的顺序，有利于预测用户将来可能的行为，从而推荐合适的资源。

Web 日志记录了用户的一系列点击流信息，而对于用户的一次请求相应地在服务器上记录多条信息且各记录之间并无直接的关联关系，因此分析 Web 日志就需要依靠数据挖掘技术从 Web 日志的点击流数据中提取有用的信息。从更广义的角度讲，数据挖掘就是从一些事实或者观察数据集合中寻找模式的决策支持过程，发现的知识可用于信息管理、查询优化、决策支持、过程控制等领域，因此数据挖掘是数据库研究中一个很有应用价值的新领域，它又是一门交叉学科，融合了数据库技术、人工智能、机器学习、神经网络、统计学等多个领域的理论和技术。

Web 挖掘是将数据挖掘方法运用于 Web 数据，提取隐藏其中的、有用的、新颖的模式或知识发现的过程[1]。Web 日志挖掘是 Web 挖掘的一个主要分支，它

旨在从大量访问者的访问历史记录中，挖掘网站的频繁使用模式、用户访问行为模式、具有相似行为的用户群等信息，使人们能够充分了解 Web 站点的使用情况和使用 Web 站点的用户行为模式，从而对 Web 站点优化组织和更好地为用户提供服务，提高 Web 站点的访问量和性能。

1.2 国内外研究现状

国外的专家学者对 Web 日志挖掘做了大量的研究，如：Cooley R, Mobasher B[2]等人首次给出 Web 挖掘的定义，并且给出一个关于 Web 访问信息挖掘的系统 WEBMINER。通过对 Web 站点的日志进行处理，将数据组织成传统的数据挖掘方法能够处理的事务数据形式，然后利用传统的数据挖掘方法（如关联规则发现算法）进行处理，所得出的挖掘结果也是传统的数据挖掘结果。Chen M S[3]等人首先将数据挖掘技术应用于 Web 服务器日志挖掘，发现用户的浏览模式。提出最大向前引用(Maximal Forward Reference，MFR)系列的概念。将用户会话分割成一系列的事务，然后采用与关联规则相似的方法挖掘频繁浏览路径。Buchner A G，Mulvenna M D [4]等人首次提出将数据挖掘技术应用于电子商务的环境下，以发现市场智能。挖掘的对象不仅包括日志、Web 页面，也包括市场数据，并且给出了在电子商务环境下挖掘的一个总框架。Zaiane[5]等人将 Web 服务器日志保存为数据立方体(Data Cube)，然后在其上执行在线数据分析处理(OLAP)的各种操作，如提升、钻取等，用于发现用户的访问模式。

国内学者在 Web 日志挖掘方面也开展了大量的工作，如：西安交通大学沈均毅教授[6]等人提出：首先以 Web 站点的 URL 为行、以 UserID 为列，建立 URL-UserID 关联矩阵，元素值为用户的访问额次数，然后，对列向量进行相似性分析得到相似客户群体，对行向量进行相似度量获得相关 Web 页面，对相关页面进行进一步处理，以发现频繁访问路径。并提出了 Web 页面和群体的模糊聚类算法。西安交通大学陆丽娜教授[7]等人，采用基于事务的方法，研究 Web 日志挖掘预处理及用户访问序列模式挖掘方法，提出了一种基于扩展有向树模型进行用户浏览模式识别的 Web 日志挖掘方法。中国科技大学王熙法教授[8]等人提出了基于神经网络的 Web 用户行为聚类分析方法，即首先对 Web 服务器的日志文件进行分析，再进行会话分析，从会话向量中找出频繁数

据集，进行归一化处理后，生成模式向量，采用 SOFM 模型进行聚类，最后生成用户聚类。中国科学院数学研究所周龙镶教授[9]等人，分析了 Web 用户浏览活动规律，提出了有关 WWW 浏览路径的一些基本概念，设计了基于用户访问模式的浏览路径优化算法。

简而言之，基于 Web 日志挖掘的研究工作大致分为以下三类[10]。

(1)以分析 Web 站点性能为目标：主要从统计学的角度，对日志数据项进行简单的统计，得到用户频繁访问页、单位时间访问数、访问数量随时间分布图等。绝大多数商用及免费的 Web 日志分析工具均属此类。

(2)以理解用户意图为目标：Chen M S 等提出的路径遍历模式(Path Traversa Pattem)的发现算法，以及 Zaiane 等使用的数据立方体方法，便是此类的典型代表。

(3)以改进 Web 站点设计为目标：通过挖掘用户的频繁访问路径和用户聚类，重构站点的页面之间的链接关系，以更适应用户的访问习惯，同时为用户提供个性化的信息服务。

同时，一些国外专业研究数据挖掘的网站已出现一些用户访问日志分析工具，主要是统计每一个页面用户访问的频次，以及用户访问页面的时间分布情况。表 1.1 是国外近几年数据挖掘研究大型项目[11]。

表 1.1　国外 Web 数据挖掘研究项目一览表

项目	应用领域	项目	应用领域
WebSIFT	普通	Web Log Miner	商业
WUM	普通	Page Gather	网站结构设计
Shahabj	普通	Manley	用户分类
Site Helper	个性化服务	Arlitt	用户分类
Letizia	个性化服务	Pitkow	用户分类
Web Watcher	个性化服务	Almeida	用户分类
Analog	商业	Shechter	提高系统效率
Web Trends	商业	Aggarwal	提高系统效率

由此可见，Web 数据挖掘在国外的发展已经比较成熟，Web 日志挖掘是一个较新的研究领域，具有广阔的发展和应用前景。然而，大多数 Web 数据挖掘

系统无法与数据库无缝集成，也不支持算法的扩展，同时由于数据挖掘标准语言的欠缺，使 Web 数据挖掘技术的应用范围仅限于领域专家。

1.3 本文内容及框架

在本文的研究中，Web 数据挖掘方面主要集中在 Web 使用挖掘的问题、Web 日志挖掘技术方面，技术实现方面主要研究的对象是规则发现，包括聚类分析算法与顺序分析和聚类分析算法。本文系统地研究了 Web 使用挖掘问题，并实现了一个 Web 日志挖掘解决方案。各章的工作是这样组织安排的：

第一章阐述了论文的研究背景及选题意义，介绍了 Web 日志挖掘国内外研究现状、研究领域和方向，及 Web 数据挖掘中存在的问题。

第二章由一般到特殊地分别介绍了数据挖掘、Web 数据挖掘。对于 Web 数据挖掘的分类、基本流程等进行了论述，形成了一个 Web 数据挖掘的较全面概述。

第三章主要对 Web 日志挖掘技术进行了研究，介绍了数据预处理的作用，提出了 Web 数据挖掘系统的数据预处理模型框架。

第四章介绍数据挖掘软件及其开发工具的发展趋势，并指出基于 SQL Server 2008 构建数据挖掘系统的优势。

第五章叙述了 SQL Server 2008 数据挖掘开发平台，并说明了利用该平台构建挖掘系统的基本步骤。

第六章介绍了 Web 日志挖掘系统的基本构架，目标和功能，及体系结构等。

第七章阐述了 Web 日志挖掘系统的设计与实现的具体过程，并介绍 Web 日志挖掘在网站建设中的作用。

第八章总结本文并展望未来的工作。